Clustering in pursuit of temporal correlation for human motion segmentation

Cheng Qian[1,2] · **Toby P. Breckon**[3] · **Zezhong Xu**[1]

Abstract Temporal correlation is an important property of the video sequence. However, most methods only accomplish the clustering of frames via the measurement of similarity between frame pair, and the temporal correlation among frames is rarely taken into account. In this paper, a method for clustering in pursuit of temporal correlation is proposed to address human motion segmentation problem. Aiming at the video sequence, a one-hot indicator vector is extracted from a frame as a frame-level feature. The description of the relationship between the features is formulated as a minimization problem with respect to a similarity graph. A temporal constraint in the form of a trace is imposed on the similarity graph to capture the temporal correlation. On the premise of the non-negative similarity graph, an optimal solution to the graph augments the relationship between the selected features and their adjacent features, while suppressing its relevance to the features that are far away from it in terms of the time span. Normalized cut is implemented on the graph so as to give clustering results. The experiments on human motion segmentation demonstrate the superior performance of the proposed method in tackling the motion data.

Keywords Human motion segmentation · Temporal correlation · Temporal clustering · Spectral clustering

✉ Cheng Qian
qc_hz@163.com

[1] Changzhou Institute of Technology, No. 666, Liaohe Road, Xinbei District, Changzhou, Jiangsu Province 213031, China

[2] Key Laboratory of Image and Video Understanding for Social Safety, Nanjing University of Science and Technology, No. 200, Xiaolingwei, Xuanwu District, Nanjing, Jiangsu Province 210094, China

[3] School of Engineering and Computing Sciences, Durham University, Durham DH1 3LE, UK

1 Introduction

Human motion segmentation divides a video sequences into a set of motion segments based on the types of actions, and obtains a wide range of applications in activity recognition, video retrieval and imitation learning [13]. At present, the accomplishment of human motion segmentation is still faced with several challenges such as the lack of well-defined activity segments and the ambiguity in motion primitives caused by the temporal variability in actions [22].

Since human motion occurs in the form of consecutive frame samples, the segmentation of the human motion can be considered as an unsupervised clustering problem concerning the frames [14, 34]. The clustering largely depends on the relationship among the frames. In particular, the similarity is crucial to the description of the relationship. As a result, some methods concentrate on the construction of a similarity graph for the frames [20, 26, 32]. In general, a graph can be established to express the similarity among the frames, and then the segmentation on the video sequence is realized through implementing the normalized cut on the graph. Benefiting from the fact that most of the frames from the same video clip have the minimal within-cluster variation, the similarity among the frames can be clearly depicted by a similarity graph, and the normalized cut on the graph thus provides an effective tool for the solution to the clustering problem.

The construction of the graph has great impact on the segmentation results. Lv et al. take the possibility of the transition from one frame to another frame as the measurement of the similarity between frame pair, and further establish a similarity graph for the frames [26]. Vogele et al. introduce a neighborhood graph into the discovery of the self-similarity structure inherent in the video sequence. The individual diagonal parts of the graph are used to determine the activities [32]. Li et al. develop a subspace clustering method for the motion segmentation [20]. The measurement of the similarity is replaced with the linear correlation among the frames. In addition, a temporal Laplacian regularization with respect to the graph is added to the solution so as to capture the order in the frames. Li et al. propose to model the spatial layout of a person with the selected joints, and the variations in Euclidean distance between joint pair during the movement constitute the graph for a motion segment [19]. Finally, the graph kernel is used to measure the similarity among motion segments, and the temporal cut points are identified via the mergers of the motion segments.

It is obvious that the clustering on the graph facilitates the efficient identification of motion segments to a large extent. However, during the construction of the graph, these graph-based methods mainly focus on the similarity among the frames. It is difficult for them to directly incorporate the temporal relationship among the frames into the graph. Since the frame samples are in nature a time series, the frames within the short temporal domain are more likely to be correlated with each other than within the long temporal domain. Moreover, in general, an entire motion primitive is only limited to a short temporal domain or an individual cycle. The frames that are far away from the current temporal domain should be excluded from the local temporal domain. As a result, most methods employ the strategy of the sliding window to handle the temporal neighborhood [10, 15]. The temporal cut points are further determined within the sliding windows. The hard cut of the time windows is prone to classifying the frames from the same motion primitive into different clusters.

In order to eliminate the error in the segmentation resulting from the lack of the temporal constraint, we try to combine the measurement of the similarity with the temporal correlation among frames under a framework. As a result, each entry of the graph not only reflects the

similarity but also measures the temporal correlation. As opposed to the sliding windows, the integration of temporal correlation into the measurement of the similarity leads to a soft cut of the time windows. The determination of the temporal cut points actually results from the tradeoff between the similarity and the temporal neighborhood. On one hand, with the constraint of the temporal neighborhood, the clustering on the graph can resist against the error in the similarities of features to some extent. On the other hand, a closed-form solution to the graph can be derived under the proposed framework, which enables the efficient representation of the relationship among features. To summarize, the contribution of the work falls into three aspects:

- A one-hot vector is extracted from each frame sample as a frame-level feature. The similarity among the frames is calculated through a bilinear similarity function with respect to the one-hot feature vectors.
- A framework for the clustering in pursuit of temporal correlation is devised. Under this framework, the temporal correlation is formulated as a constraint term in the form of trace and is enforced on the similarity function.
- In the proposed framework, a graph is derived via the minimization of an objective function with respect to it. In particular, the graph is given in the form of a closed-form solution to the objective function. The normalized cut is implemented on the graph so as to yield the clustering results, which are equivalent to the segmentation results.

The reminder of the paper is organized as follows. The related work is summarized in section 2. In section 3, the methods for clustering in pursuit of temporal correlation are presented. The experimental results are reported in section 4 with the conclusion in section 5.

2 Related work

There has been much work in the field of human motion segmentation, and a large number of attempts are devoted to solving the problem. The human motion segmentation can be regarded as a frame labeling problem. Among them, Fod et al. proposed to split the arm motion sequence into video clips at the zero-velocity crossing points [8]. A gesture was defined by the segmentation of the trajectory data with a set of threshold values [5]. Lin et al. casted the motion segmentation as a binary classification problem. Each data point belongs to either a segment point or a non-segment point [23]. Tao et al. devised a maximum correntropy based image labeling framework for the measurement of the similarity between training images and test images. According to the similarity, a label is assigned to a test image [28]. Furthermore, Tao et al. proposed a method for large sparse cone non-negative matrix factorization to learn the semantic parts of an image. The resemblances of the semantic parts are combined into a score for a label [29].

How to measure the similarity between the frame pair plays an important part in segmentation. Aoki et al. modeled the motion primitives as a Hidden Markov model (HMM), and then it was employed to recognize motion segments [1]. Zhou et al. created a generalized dynamic time alignment kernel as the measurement of the similarity between motion segment pair, and it is integrated into the kernel *k*-means framework in order to determine the cluster centers. The assignment of each motion segment to a cluster interprets the composition of a long motion [35]. Lan et al. utilized the hierarchical clustering to extract the key poses. Most of motions are

considered to be composed of these key poses, and they are grouped into the motion primitives by the latent Dirichlet allocation. The similarity between a segment and a key pose is measured to detect the segment points [18].

The information about the existing motion segments is beneficial for the segmentation. Lin et al. roughly identify the motion segments by the velocity features. The HMM is further utilized to reduce over-segmentation and refine the segmentation [21]. A support vector machine (SVM) is trained with a group of actions, and then it is used to label the windows of data. Based on the labeled sequences, a HMM created the motion templates and then outputs the segment points [31]. Avgerinakis et al. extract the motion boundary activity area for coarse localization of the activities, and further determine the precise activity boundary via sequential statistical change detection [2].

The latent structure of the distribution of features is an important property of frames originating from the same motion. Kruger et al. used region growing to split the video sequence into distinct activities and then searched for the motion primitives by means of the neighborhood graph-based similarity [16]. Gong et al. put forward the kernelized temporal cut for temporal segmentation [9]. Baptisra et al. combined a group of linear state-space models into a nonlinear system to represent the human motion. The temporal cut points are acquired via the solution to the model [4]. Recently, the motion data is assumed to reside on a union of multiple subspaces. If every cluster of frames can be accommodated with a subspace, the human motion segmentation can be transformed into the subspace segmentation [20], and subspace clustering methods are able to offer the solutions to segmentation [6, 7, 12, 24]. Especially, some subspace clustering methods were extended to deal with the time series. Among them, Tierney et al. proposed to enforce a neighborhood penalty on the sparse linear representation [30]. However, the detection of the difference between data pair is limited to a pair of data next to each other, and it actually plays a role of one order derivative. It is equivalent to smoothing over the linear representation in the temporal domain at the cost of insensitivity to the drastic change in the data. Bahadori et al. introduced the warping-based alignment deformation operators into the assignment of the data to the subspace [3]. Instead of the temporal relationship, this method largely focuses on clustering the data in the case of the deformation that happens to the time series. Tao et al. designed a method for nonnegative matrix factorization under manifold regularization to explore the neighborhood relationship among the image regions [27]. Aiming at the human motion problem, we try to exploit the temporal neighborhood inherent in frames to address the motion segmentation problem under the graph-based framework.

3 Proposed method

As shown in Fig. 1, our proposed method is composed of two components: (a) Frame-level feature extraction and (b) similarity graph-based clustering. While the former part is in charge of the description of frames, the latter part establishes the similarity graph for all frames, and the motion segmentation is realized via the clustering on the similarity graph.

3.1 Frame-level feature for human motion

For a video sequence containing human motion, the features should be constructed as the descriptors for frames. Here, a set of temporal words are employed to encode

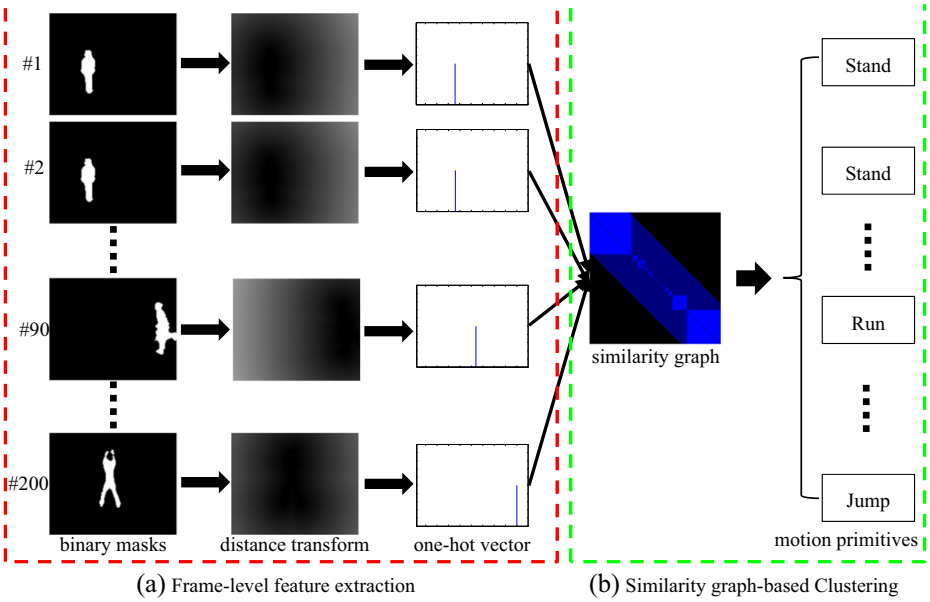


Fig. 1 Flowchart of the proposed method. **a** The binary masks are extracted from all frames. The Euclidean distance transform is then implemented on the masks. *k*-means clustering on the results of the Euclidean distance transform produces the one-hot vectors, which act as the frame-level features. This part is denoted by a red bounding box. **b** The similarity graph is established based on the products of the frame-level features. The clustering on the similarity graph yields the motion primitives. This part is denoted by a green bounding box

frames. First of all, a binary mask is extracted from each frame, and then the Euclidean distance transform is implemented on these binary masks [11]. Considering the representative ability and the efficiency, *k*-means clustering is taken as one-hot encoder. Resorting to *k*-means clustering, the results generated from the Euclidean distance transform can further fall into a group of clusters. These clusters constitute the temporal words. Each frame is assigned to a cluster, which is equivalent to being encoded with the temporal words. A binary indicator vector labels the assignment. As a result, these vectors in the form of the one-hot vectors serve as frame-level features.

The entire procedure of frame-level feature extraction is illustrated by the red bounding box in Fig. 1. It can be seen that it is mainly made up of three steps: extraction of binary masks, the Euclidean distance transform over the masks and the generation of the one-hot indicator vector. Note that each element of the one-hot vector is either 0 or 1. In detail, a label 1 indicates that the result is classified into the current cluster, while 0 represents a label that the result is out of the current cluster. Therefore, it holds true that all frame-level features in the form of the one-hot vectors are non-negative.

3.2 Measure of similarity

In general, it is taken for granted that the similarity between a pair of frame-level features represents the correlation. A pairwise similarity function with respect to features can give the measurement of similarity. Since a similarity function in the bilinear form offers an effective

metric for the similarity [17], the measurement of similarity can be defined as a pairwise function s as follows:

$$s(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \cdot \mathbf{W} \cdot \mathbf{x}_j, \quad \mathbf{x}_i, \quad \mathbf{x}_j \in \mathbb{R}^{d \times 1} \tag{1}$$

Where $\mathbf{x}_i$ and $\mathbf{x}_j$ are two arbitrary d-dimensional features. $\mathbf{W} \in \mathbb{R}^{d \times d}$ is a similarity metric matrix. Considering that all of the frame-level features are the non-negative binary vectors, it is reasonable for $\mathbf{W}$ to be an identity matrix $\mathbf{I}$. As a result, the function s in Eq. (1) can be further reduced to a simpler form.

$$s(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \cdot \mathbf{I} \cdot \boldsymbol{x}_j = \boldsymbol{x}_i^T \boldsymbol{x}_j \tag{2}$$

Suppose that the features $\mathbf{X} = [\mathbf{x}_t, \mathbf{x}_{t+1}, \cdots, \mathbf{x}_{t+n-1}]$ from t-th frame to $(t+n-1)$-th frame are gathered, the similarity matrix $\mathbf{Z}$ can be readily obtained by computing as Eq. (2) on all pairs of features. Consequently, the similarity among the features can be expressed in the form of a matrix.

$$\mathbf{Z} = \mathbf{X}^T \mathbf{X}, \quad \mathbf{X} \geq \mathbf{0} \tag{3}$$

Due to the non-negative matrix $\mathbf{X}$, $\mathbf{Z} \geq 0$ holds. Here, it is noticeable that $\mathbf{Z} \in \mathbb{R}^{n \times n}$ constructs a similarity graph that describes the similarity among features. It is illustrated by the green bounding box in Fig. 1.

3.3 Combination of similarity and temporal correlation

In Eq. (3), it is obvious that the graph $\mathbf{Z}$ only takes the similarity of features into consideration, whereas the temporal correlation is ignored. It is expected that $\mathbf{Z}$ not only reflects the similarity but also incorporates the temporal correlation into itself. Hence, $\mathbf{Z}$ should be reconstructed so as to satisfy the requirement of the combination of the similarity and the temporal correlation.

According to the above requirement of reconstruction, it is essential that a framework for the combination should be devised. Under the framework, on one hand, the similarity among the features can be sufficiently approximated by $\mathbf{Z}$. On the other hand, the temporal correlation among features can be captured through imposing a constraint on $\mathbf{Z}$. Therefore, the requirement can be formulated as an objective function $J(\mathbf{Z})$ with respect to $\mathbf{Z}$ as follows:

$$J(\mathbf{Z}) = \left\| \mathbf{Z} - \mathbf{X}^T \mathbf{X} \right\|_2^2 + \lambda \cdot \varphi(\mathbf{Z}), \quad s.t. \quad \mathbf{Z} \geq \mathbf{0} \tag{4}$$

Where the first term is an approximation term in the form of ℓ_2 norm, and the second term $\varphi(\mathbf{Z})$ denotes a temporal constraint. $\lambda > 0$ is a tradeoff parameter. As a result, the obtainment of $\mathbf{Z}$ is converted into a problem defined by the objective function $J(\mathbf{Z})$.

3.4 Temporal correlation

In terms of the time span, the adjacent frames often exhibit stronger correlation than the frames with great temporal distance. This defines the concept of temporal neighborhood within the sequence. In our method, the temporal correlation of the features largely concentrates on the temporal neighborhood. Recalling the temporal constraint $\varphi(\mathbf{Z})$ in Eq. (4), it is required that the constraint is able to act as a time window that confines the measurement of similarity to a limited time interval. For the purpose of the temporal confinement, a simple yet effective way to design the temporal constraint is proposed here.

$$\varphi(\mathbf{Z}) = \mathrm{tr}(\mathbf{A}\mathbf{Z}) \tag{5}$$

Where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a weight matrix that plays a role as a filter in the temporal domain. The trace of $\mathbf{AZ}$ is actually a summation of the products between the row vectors of $\mathbf{A}$ and the column vectors of $\mathbf{Z}$.

$$\mathrm{tr}(\mathbf{AZ}) = trace\left(\begin{bmatrix} \boldsymbol{A}_1\boldsymbol{Z}_1 & \cdots & \boldsymbol{A}_1\boldsymbol{Z}_n \\ \boldsymbol{A}_2\boldsymbol{Z}_1 & & \boldsymbol{A}_2\boldsymbol{Z}_n \\ \vdots & \ddots & \vdots \\ \boldsymbol{A}_n\boldsymbol{Z}_1 & \cdots & \boldsymbol{A}_n\boldsymbol{Z}_n \end{bmatrix}\right) = \sum_{i=1}^{n} \boldsymbol{A}_i\boldsymbol{Z}_i \tag{6}$$

Where $\mathbf{A}_i$ denotes the i-th row vector of $\mathbf{A}$, and $\mathbf{Z}_i$ denotes the i-th column vector of $\mathbf{Z}$ respectively.

The goal of capturing the temporal neighborhood can be reached by designing an appropriate weight matrix $\mathbf{A}$. Hence, in detail, each entry in $\mathbf{A}$ should meet the requirement as follows:

$$\begin{cases} \mathbf{A}_{i,k} = -1, & i-\tau/2 \le k \le i+\tau/2 \\ \mathbf{A}_{i,k} = 1, & otherwise \end{cases}, \quad s.t.\ \ i,k = 1,2,\cdots,n \tag{7}$$

Where $\tau > 0$ controls the length of the time interval. As a result, the trace in Eq. (6) can be further expanded in the form of elements as follows:

$$\mathrm{tr}(\mathbf{AZ}) = \sum_{i=1}^{n} \boldsymbol{A}_i\boldsymbol{Z}_i = \sum_{i=1}^{n} \left(\sum_{t=1}^{i-\tau/2-1} \boldsymbol{Z}_{t,i} + \sum_{t=i+\tau/2+1}^{n} \boldsymbol{Z}_{t,i} - \sum_{t=i-\tau/2}^{i+\tau/2} \boldsymbol{Z}_{t,i} \right) \tag{8}$$

Such setting of the temporal constraint paves the way for the subsequent temporal confinement. In the case of non-negative $\mathbf{Z}$, the minimization of Eq. (8) is equivalent to augmenting the correlation with the adjacent features and suppressing the relevance to the distant features simultaneously.

Both the minimization of the approximation error and the pursuit of the temporal correlation can be accomplished simultaneously via the minimization of the objective function $J(\mathbf{Z})$.

$$\mathbf{Z} = \mathrm{argmin}_{\mathbf{Z}} J(\mathbf{Z}), \quad s.t.\ \ \mathbf{Z} \ge \mathbf{0} \tag{9}$$

Hence, the calculation of the matrix $\mathbf{Z}$, which combines the similarity with the temporal correlation, turns out to be an optimization problem. In detail, it is expanded as follows.

$$\mathbf{Z} = \mathrm{argmin}_{\mathbf{Z}} \left(\left\| \mathbf{Z} - \mathbf{X}^T\mathbf{X} \right\|_2^2 + \lambda \cdot \mathrm{tr}(\mathbf{AZ}) \right), \quad s.t.\ \ \mathbf{Z} \ge \mathbf{0} \tag{10}$$

It is evident that the objective function $J(\mathbf{Z})$ is a convex function. The derivative of $J(\mathbf{Z})$ with respect to $\mathbf{Z}$ forms a constraint formula for the solution.

$$\frac{dJ(\mathbf{Z})}{\mathbf{dZ}} = \mathbf{Z} - \mathbf{X}^T\mathbf{X} + \frac{\lambda}{2} \cdot \mathbf{A} = \mathbf{0}, \quad s.t. \cdot \mathbf{Z} \ge \mathbf{0} \tag{11}$$

The solution to Eq. (11) yields $\mathbf{Z}^*$.

$$\mathbf{Z}^* = \max\left(\mathbf{X}^T\mathbf{X} - \frac{\lambda}{2} \cdot \mathbf{A}, 0 \right) \tag{12}$$

In this way, the solution $\mathbf{Z}^*$ guarantees the integration of the temporal correlation into the measurement of the similarity.

To investigate the effect of the temporal correlation on the similarity matrix, the proposed approach is conducted on a group of frame-level features to obtain a tentative observation. In detail, three video clips, each of which contains 498 frames, are drawn from Keck dataset.[1] They are concatenated into a video sequence. The similarity matrix $\mathbf{X}^T\mathbf{X}$ is shown in Fig. 2(a), while its counterpart $\mathbf{Z}^*$ calculated as Eq. (12) is shown in Fig. 2 (b).

As for the similarity matrix displayed in Fig. 2(a), it can be seen that not only the frames belonging to the same video clip correlate with each other, but also the frames from different video clips are strongly relevant to each other. The clustering easily leads to the ambiguity in segmentation. According to Fig. 2(b), there only exists the correlation between frame pair along the diagonal line. The temporal constraint enforces the time windows for the pursuit of the correlation to be shrunk down to a local interval surrounding each frame. Hence, the relationship between a pair of frames from different video clips can be eliminated, thereby diminishing the false correlations.

3.5 Normalized cut on similarity matrix

From the perspective of graph, $\mathbf{Z}^*$ actually acts as an affinity graph. Therefore, once that the similarity graph $\mathbf{Z}^*$ is derived, it is straightforward for normalized cut to be implemented on $\mathbf{Z}^*$ to get the clustering result.

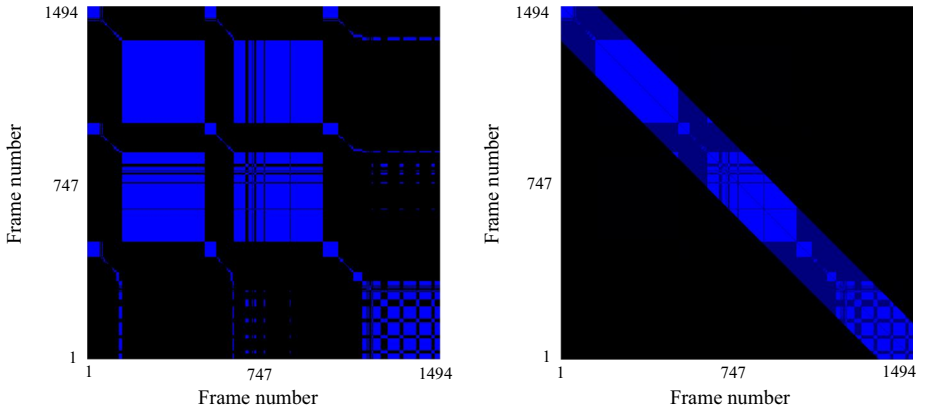
4 Experiment and analysis

To verify the performance, the proposed method is carried out on the video sequences containing human motion. At first, we elaborate on the dataset, the parameter setting and the criteria for evaluation. Subsequently, the experimental results are presented and discussed. Finally, to further evaluate the performance, our method is compared with several state-of-the-art graph-based clustering methods, including sparse subspace clustering (SSC) [7], low rank representation (LRR) [24], local subspace analysis (LSA) [33], least square regression (LSR) [25], ordered subspace clustering (OSC) [30], aligned cluster analysis [3] and hierarchical aligned cluster analysis [17].

4.1 Dataset and implementation detail

As for the dataset, Keck action recognition dataset and Weizmann dataset[2] are chosen to serve as the test data. In detail, Keck dataset consists of the actions of three subjects. Each subject performs 14 actions. In Fig. 3, three frames are drawn from three video clips respectively, each of which exhibits a gesture different from the others. Weizmann dataset comprises the video sequences of 9 subjects. Among them, each man finishes 10 actions. As a public dataset, it is available for motion segmentation and action recognition. Likewise, Fig. 4 shows three frames from three different video clips. These frames represent the components of different actions. To construct the data for test, a collection of the action video clips are selected from the dataset at random, and they are concatenated into a long video sequence. The human motion segmentation is implemented on these long video sequences to discriminate each video clip against the others.

[1] http://www.umiacs.umd.edu/~zhuolin/Keckgesturedataset.html
[2] http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html



(a) Similarity matrix (b) Similarity matrix with temporal constraint

Fig. 2 **a** The similarity matrix. **b** The similarity matrix with temporal constraint. Each entry of the matrix is represented with the color to indicate its intensity. The stronger the correlation between feature pair is, the lighter blue the entry is represented with, and vice versus

As far as the parameter setting is concerned, the dimension of the one-hot vector is 50. It is equivalent to the number of clusters involved in the k-means. The tradeoff parameter λ is set to be 1.9 in Eq. (4). These parameters are fixed throughout all experiments. In addition, the length of time interval τ in Eq. (7) has significant influence on the segmentation. As a result, the influence of τ on the performance is investigated. In order to better evaluate the quality of human motion segmentation, four criteria including the purity, the precision, the recall and F-measure, are introduced into the evaluation.

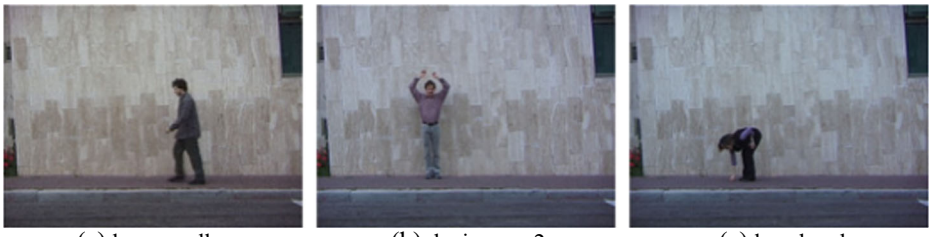
4.2 Experimental results

Keck dataset The proposed method is evaluated against the other state-of-the-art work on Keck dataset. As Fig. 5 shows, it is evident that our method achieves better results than the other methods in terms of purity, precision and F-measure. In particular, in contrast to most methods, our method achieves at least 20% improvement in purity. The advantage of the proposed method over the other methods mainly results from the introduction of the temporal correlation into the graph. The graph lays the foundation for the identification of the temporal cut points.

In addition, since the performance of segmentation varies with the number of clusters, a group of video sequences consisting of different number of short video clips are taken as the test data. In Fig. 5, a quantitative evaluation of the performance with respect to the number of

(a) 1st action (b) 5th action (c) 10th action

Fig. 3 Three frames are drawn from the video clips of Keck dataset. These video clips contain different actions



(a) lyova walk (b) denis wave2 (c) lena bend

Fig. 4 Three frames are drawn from the video clips of Weizmann dataset. These video clips contain different actions of different men

clusters is also given. It can be seen that, as the number of clusters grows, the performances of all methods deteriorate. Especially, as the cluster number exceeds 6, the performance of our method drops dramatically. This is mainly attributed to that the mixture of more motion primitives tends to cause more ambiguity in the segmentation. Besides, the symmetry gestures also make it more difficult to segment the actions. However, despite the variation in the number of clusters, our method can still maintain the advantage over the other methods in terms of the purity, the precision as well as F-measure.

Weizmann dataset To further validate the clustering performance, our method is also conducted on Weizmann dataset. Compared with Keck dataset, the actions in Weizmann dataset is more complicated. Fig. 6 illustrates the quantitative measurements of the clustering performances on Weizmann dataset. As for the measurement of the purity, our method outperforms the other methods by a good margin. From the perspective of precision shown in Fig. 6(b), the ambiguities in the clustering results generated by the other methods are more severe than our method. However, in Fig. 6(c), it is also noted that, both ACA and HACA are able to achieve comparable higher recall

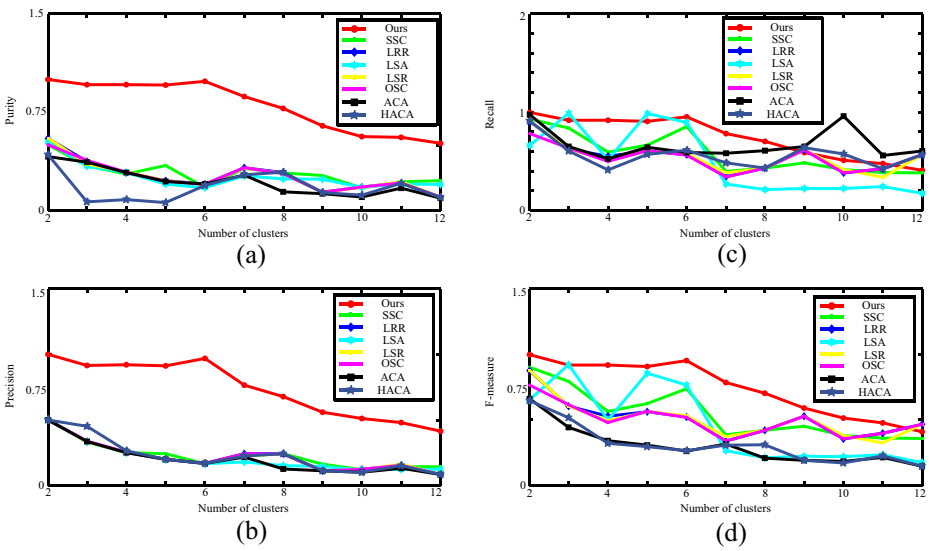


Fig. 5 A quantitative evaluation of the performance on Keck dataset. **a** The purity varies with the number of clusters. **b** The precision varies with the number of clusters. **c** The recall varies with the number of clusters. **d** F-measure varies with the number of clusters

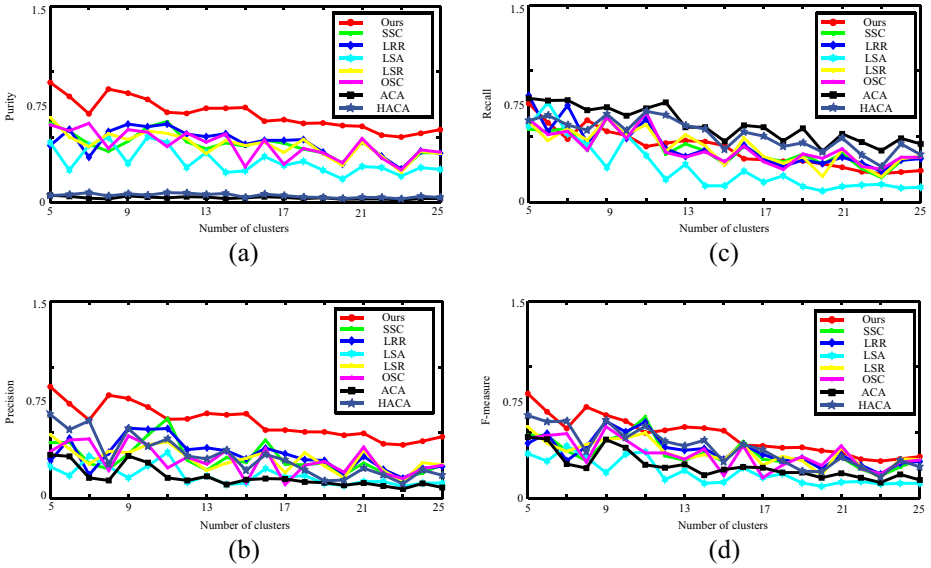


Fig. 6 A quantitative evaluation of the performance on Weizmann dataset. **a** The purity varies with the number of clusters. **b** The precision varies with the number of clusters. **c** The recall varies with the number of clusters. **d** F-measure varies with the number of clusters

than our method. It implies that the proposed method builds up the clusters with more reliable frames, but at the cost of excluding some frames from its true belonging.

As for the influence of the number of clusters on the performance, it is observed that, as the number of clusters grows, the tendency of the performance towards degradation also happens on Weizmann dataset. Nevertheless, like Keck dataset, most of the time, our method still performs better than the other methods on Weizmann dataset.

4.3 Temporal constraint

To depict the relationship of an arbitrary data pair sampled from the data stream, it is important that the measurement of relationship strike a balance between the similarity and the temporal correlation. Under the proposed framework, the temporal correlation in the form of the temporal neighborhood confines the measurement of the similarity to a local temporal domain. It only allows the features in the proximity of the current feature to join in the measurement. Meanwhile, the features that are distant from it are excluded from the measurement. To gain more insights on the influence of temporal constraint on the segmentation performance, the temporal constraint term is separated from the objective function in Eq. (9). It is then tested on Keck dataset and Weizmann dataset respectively in order to give the quantitative evaluations.

In Table 1, the quantitative comparison of segmentation performances on Keck dataset is given. It can be seen that, no matter how many motion primitives need to be segmented, our method outperforms the method without the temporal constraint by a large margin.

According to the segmentation results on Weizmann dataset listed in Table 2, in terms of the accuracy, our method performs better than the method without the temporal correlation. However, it also deserves attention that, as the number of motion primitives grows, the method without the temporal constraint is close to or even exceeds our method in terms of recall. Due

Table 1 A quantitative comparison of segmentation performance on Keck dataset between the clustering method without temporal constraint and our proposed method. In comparison, the better performance is indicated with the bold font. (without: the clustering method without temporal constraint, with: the proposed method with temporal constraint)

Metric	Cluster = 2		Cluster = 4		Cluster = 6		Cluster = 8	
	without	with	without	with	without	with	without	with
Purity	0.54	**1.00**	0.30	**0.96**	0.21	**0.98**	0.28	**0.78**
Precision	0.51	**1.00**	0.25	**0.92**	0.17	**0.97**	0.23	**0.68**
Recall	0.69	**1.00**	0.56	**0.92**	0.60	**0.95**	0.45	**0.70**
F-measure	0.59	**1.00**	0.35	**0.92**	0.26	**0.96**	0.30	**0.69**

to the constraint on the temporal coverage of features, in general, the number of features that needs to be clustered is less than the method without the temporal constraint. In fact, to ensure the purity of clustering, the frames with low confidences are often assigned to another cluster that is not consistent with their true cluster. As a result, it is inevitable for our method to incur the loss of discarding some frames from the same cluster. However, a delicate selection of time interval can alleviate the loss to some extent.

4.4 Length of time interval

Since the length of time interval τ plays an important role in capturing temporal correlation, it is necessary to choose an appropriate τ for the time interval. Based on the experiments on Keck dataset and Weizmann dataset, the influence of the change in τ on the clustering performance is empirically observed. The corresponding experimental results are shown in Fig. 7. Note that, instead of the concrete setting of τ, the length ratio τ/n is taken as a measurement of coverage of temporal neighborhood. According to the experimental results, as far as the clustering performances are concerned, the short time interval is more competent to pursue the temporal correlation than the long time interval. This is largely attributed to that the features gathered in a local temporal domain are more prone to being consistent with each other. The high coherence of the neighboring features allows for the assignments of these frames to the same cluster.

In addition, when the number of clusters changes, the influence of τ on the segmentation is also different from each other. When the length ratio is small, it is obvious that the accuracy of

Table 2 A quantitative comparison of segmentation performances on Weizmann dataset between the clustering method without temporal constraint and our method. In comparison, each better performance is indicated with the bold font. (without: the clustering method without temporal constraint, with: the proposed method with temporal constraint)

Metric	Cluster = 5		Cluster = 11		Cluster = 17		Cluster = 23	
	without	with	without	with	without	with	without	with
Purity	0.54	**0.91**	0.44	**0.69**	0.29	**0.63**	0.24	**0.50**
Precision	0.35	**0.85**	0.26	**0.60**	0.10	**0.52**	0.13	**0.40**
Recall	0.52	**0.75**	0.39	**0.42**	**0.39**	0.32	**0.25**	0.23
F-measure	0.42	**0.80**	0.31	**0.49**	0.16	**0.40**	0.17	**0.29**

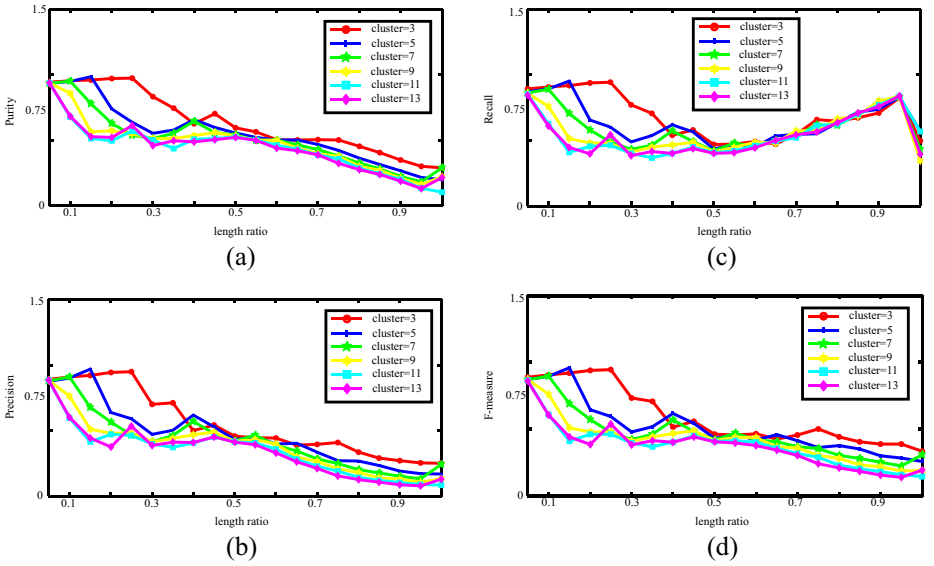


Fig. 7 The measurement of the influence of the time interval on human motion segmentation when facing different number of motion primitives from Keck dataset. **a** The purity versus length ratio. **b** The precision versus length ratio. **c** The recall versus length ratio. **d** F-measure versus length ratio

the segmentation on the fewer actions is higher than more actions. However, as Fig. 8 shows, the extension of the time interval bridges the gap between the performances on different number of clusters. Hence, compared with the long time interval, the selection of the short time interval is in favor of the pursuit of temporal correlation.

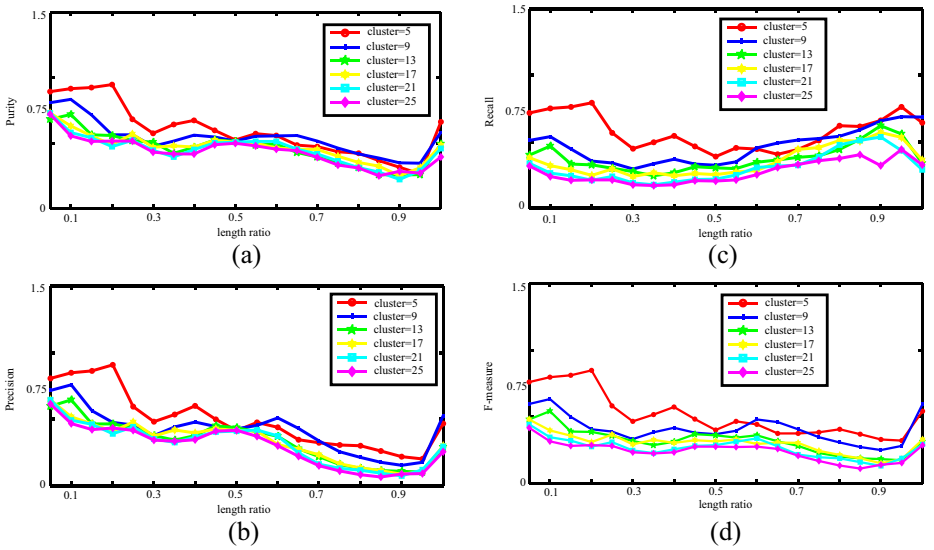


Fig. 8 The measurement of the influence of the time interval on human motion segmentation when facing different number of motion primitives from Weizmann dataset. **a** The purity versus length ratio. **b** The precision versus length ratio. **c** The recall versus length ratio. **d** F-measure versus length ratio

Table 3 The performance of the motion segmentation on Keck dataset in the case of different setting of the tradeoff parameter λ. (For the simplicity of presentation, "puri" stands for purity, "prec" stands for precision, "reca" stands for recall, "Feam" stands for F-measure)

	Cluster = 5				Cluster = 9				Cluster = 13			
	puri	prec	reca	Fmea	puri	prec	reca	Fmea	puri	prec	reca	Fmea
λ=0.4	0.37	0.38	0.65	**0.63**	0.53	0.47	0.54	**0.54**	0.57	0.50	0.53	**0.52**
λ=0.9	0.52	0.49	0.55	**0.63**	0.45	0.41	0.54	**0.54**	0.57	0.51	0.53	**0.52**
λ=1.4	0.59	0.53	0.60	**0.63**	0.50	0.50	0.62	**0.54**	0.63	0.53	0.56	**0.52**
λ=1.9	0.59	0.53	0.60	**0.63**	0.67	0.60	0.64	**0.54**	0.56	0.50	0.53	**0.52**
λ=2.4	0.59	0.53	0.60	**0.63**	0.68	0.62	0.64	**0.54**	0.59	0.48	0.50	**0.52**
λ=2.9	0.61	0.54	0.61	**0.63**	0.63	0.55	0.57	**0.54**	0.51	0.45	0.47	**0.52**
λ=3.4	0.59	0.52	0.59	**0.63**	0.59	0.54	0.59	**0.54**	0.59	0.49	0.52	**0.52**
λ=3.9	0.61	0.53	0.56	**0.63**	0.63	0.54	0.54	**0.54**	0.46	0.39	0.42	**0.52**

4.5 The determination of tradeoff parameter

In Eq. (10), it can be seen that the tradeoff parameter λ actually strikes a balance between the similarity among frame-level features and temporal correlation. λ exerts the influence on the motion segmentation. For the sake of the determination of λ, the performance of the segmentation in the case of different setting of λ is evaluated, and the corresponding results are reported in Table 3 and Table 4 respectively. Although the criteria such as purity, precision and recall fluctuate with the change of λ, it is noticeable that, regardless of the dataset, F-measure always keeps constant. Considering that F-measure combines precision and recall, it reflects the overall assessment of the performance. Hence, when λ ranges from 0.4 to 3.9, the different setting of λ does not have considerable impact on segmentation. Besides it, as the number of clusters increases, it can be observed that, for an arbitrary setting of λ, the performance always degrades. Hence, λ is not a key factor for the segmentation performance.

5 Conclusion

A method for clustering in pursuit of temporal correlation inherent in the video sequence is proposed to tackle the human motion segmentation. A one-hot vector is utilized to characterize a frame as the frame-level feature. A framework is established to measure the relationship between an arbitrary pair

Table 4 The performance of the motion segmentation on Weizmann dataset in the case of different setting of the tradeoff parameter λ. (For the simplicity of presentation, "puri" stands for purity, "prec" stands for precision, "reca" stands for recall, "Feam" stands for F-measure)

	Cluster = 12				Cluster = 18				Cluster = 24			
	puri	prec	reca	Fmea	puri	prec	reca	Fmea	puri	prec	reca	Fmea
λ=0.4	0.89	0.81	0.16	**0.16**	0.89	0.81	0.12	**0.13**	0.78	0.72	0.11	**0.11**
λ=0.9	0.90	0.84	0.18	**0.16**	0.88	0.81	0.11	**0.13**	0.84	0.81	0.11	**0.11**
λ=1.4	0.93	0.88	0.16	**0.16**	0.88	0.83	0.11	**0.13**	0.83	0.82	0.11	**0.11**
λ=1.9	0.90	0.85	0.15	**0.16**	0.91	0.87	0.11	**0.13**	0.80	0.78	0.11	**0.11**
λ=2.4	0.90	0.83	0.15	**0.16**	0.93	0.89	0.11	**0.13**	0.84	0.77	0.10	**0.11**
λ=2.9	0.97	0.95	0.16	**0.16**	0.94	0.90	0.11	**0.13**	0.77	0.71	0.09	**0.11**
λ=3.4	0.89	0.83	0.15	**0.16**	0.94	0.90	0.11	**0.13**	0.84	0.78	0.10	**0.11**
λ=3.9	0.90	0.85	0.16	**0.16**	0.92	0.89	0.11	**0.13**	0.81	0.77	0.10	**0.11**

of features, including the similarity and the temporal neighborhood. In this framework, the temporal correlation is formulated as a constraint term in the form of trace with respect to the weight matrix. It equips the similarity matrix with the ability to represent the time series. As a result, the measurement of the relationship is transformed into an optimization problem with respect to the similarity graph. The solution to the problem produces a graph, into which the temporal correlation is integrated. The segmentation is accomplished via the normalized cut on the graph. Finally, the experiments on the human motion segmentation validate the effectiveness of the proposed methods.

As for the future work, we will seek for a more effective approach to automatically determine the length of time interval so as to further enhance the accuracy of human motion segmentation.

Acknowledgments This work has been partly supported by the National Natural Science Foundation of China (Grant No. 61602063), the Project of Natural Science Research of Higher Education Institutions of Jiangsu Province (Grant No. 15KJB520003) and the Project supported by the Jiangsu Key Laboratory of Image and Video Understanding for Social Safety (Nanjing University of Science and Technology) (Grant No. SHAQKFKT201505). It is also sponsored by Qing Lan Project.

References

1. Aoki T, Venture G, Kulic D (2013) Segmentation of human body movement using inertial measurement unit. In: Systems, Man, and Cybernetics (SMC), 2013 I.E. International Conference on. pp. 1181–1186. doi: https://doi.org/10.1109/SMC.2013.205
2. Avgerinakis K, Briassouli A, Kompatsiaris Y (2016) Activity detection using sequential statistical boundary detection (SSBD). Comput Vis Image Underst 144:46–61
3. Bahadori Moammad T, Kale D, Fan Y, Liu Y (2015) Functional subspace clustering with application to time series. In: Proceedings of the 32nd International Conference on Machine Learning. pp. 228–237
4. Baptisra R, Bo A, Hayashibe M (2017) Automatic human movement assessment with switching linear dynamic system: motion segmentation and motor performance. IEEE Trans Neural Syst Rehabi Eng 25(6):628–640
5. Beh J, Han D, Ko H (2014) Rule-based trajectory segmentation for modeling hand motion trajectory. Pattern Recogn 47(4):1586–1601
6. Chen J, Yang J (2014) Robust subspace segmentation via low-rank representation. IEEE Trans Cybern 44(8):1432–1445
7. Elhamifar E, Vidal R (2013) Sparse subspace clustering: algorithm, theory, and applications. IEEE Trans Pattern Anal Mach Intell 35(11):2767–2781
8. Fod A, Mataric Maja J, Jenkins Odest C (2012) Automated derivation of primitives for movement classification. Auton Robot 12(1):39–54
9. Gong D, Medioni ZX (2014) Structured time series analysis for human action segmentation and recognition. IEEE Trans Pattern Anal Mach Intell 36(7):1414–1427
10. Gong D, Medioni G, Zhu S, Zhao X (2012) Kernelized temporal cut for online temporal segmentation and recognition. In: 12th European Conference on Computer Vision. pp 229–243
11. Gorelick L, Blank M, Shechtman E, Irani M, Basri R (2007) Actions as space-time shapes. IEEE Trans Pattern Anal Mach Intell 29(12):2247–2253
12. Hu H, Feng J, Zhou J (2015) Exploiting unsupervised and supervised constraints for subspace clustering. IEEE Trans Pattern Anal Mach Intell 37(8):1542–1557
13. Hwang Kao S, Jiang Wei C, Chen Yu J, Shi H (2017) Motion segmentation and balancing for a biped robot's imitation learning. IEEE Trans Ind Informat 13(3):1099–1108
14. Jung H, Hong K (2017) Modeling temporal structure of complex actions using bag-of-sequencelets. Pattern Recogn Lett 85:21–28
15. Kruger B, Vogele A, Willig T, Yao A, Klein R, Weber A (2017) Efficient unsupervised temporal segmentation of motion data. IEEE Trans Pattern Anal Mach Intell 19(4):787–812
16. Kruger B, Vogele A, Willig T, Yao A, Klein R, Weber A (2017) Efficient unsupervised temporal segmentation of motion data. IEEE Trans Multimedia 19(4):797–812
17. Kulis B (2013) Metric learning: a survey. Foundations Trends Mach Learn 5(4):287–364
18. Lan R, Sun H (2015) Automated human motion segmentation via motion regularities. Vis Comput 31(1):35–53
19. Li M, Leung H (2016) Graph-based representation learning for automatic human motion segmentation. Multimed Tools Appl 75(15):9205–9224

20. Li S, Li K, Fu Y (2015) Temporal subspace clustering for human motion segmentation. In: Computer Vision (ICCV), 2015 I.E. International Conference on. pp. 4453–4461. doi:https://doi.org/10.1109/ICCV.2015.506
21. Lin Jonathan F, Kulic D (2014) On-line segmentation of human motion for automated rehabilitation exercise analysis. IEEE Trans Neural Syst Rehabil Eng 22(1):168–180
22. Lin Jonathan F, Karg M, Kulic D (2016) Movement primitive segmentation for human motion modeling: a framework for analysis. IEEE Trans Hum-Mach Syst 46(3):325–339
23. Lin F J, Joukov V, Kulic D (2014) Human motion segmentation by data point classification. In: Engineering in Medicine and Biology Society 36th Annual International Conference of the IEEE. pp. 9–13
24. Liu G, Lin Z, Yu Y (2010) Robust subspace segmentation by low-rank representation. In: The 27th International Conference on Machine Learning. pp. 663–670
25. Lu C, Min H, Zhao Z, Zhu L, Huang D, Yan S (2012) Robust and efficient subspace segmentation via least squares regression. In: 12th European Conference on Computer Vision. pp. 347–360
26. Lv N, Feng Z, Zhao X (2013) Human motion capture data segmentation based on graph partition. In: Image and Signal Processing (CISP), 2013 6th International Congress on. pp. 1117–1121. doi: https://doi.org/10.1109/CISP.2013.6745223
27. Tao D, Cheng J, Song M, Lin X (2016) Manifold ranking-based matrix factorization for saliency detection. IEEE Trans Neural Netw Learn Syst 27(6):1122–1134
28. Tao D, Cheng J, Gao X, Li X, Deng C (2017) Robust sparse coding for mobile image labeling on the cloud. IEEE Trans Circuits Syst Video Technol 27(1):62–72
29. Tao D, Tao D, Li X, Gao X (2017) Large sparse cone non-negative matrix factorization for image annotation. ACM Trans Intell Syst Technol 8(3):1–21
30. Tierney S, Gao J, Guo Y, Guo Y (2014) Subspace clustering for sequential data. In: Computer Vision and Pattern Recognition (CVPR), 2014 I.E. Conference on. pp. 1019–1026. doi:https://doi.org/10.1109/CVPR.2014.134
31. Vicente I, Kyrki V, Kragic D, Larsson M (2012) Action recognition and understanding through motor primitives. Adv Robot 21(15):1687–1707
32. Volgele A, Kruger B, Klein R (2014) Efficient unsupervised temporal segmentation of human motion. In: Proceedings of the 2014 ACM SIGGRAPH/Eurographics Symposium on Computer Animation. pp. 167–176. doi:https://doi.org/10.2312/sca.20141135
33. Yan J, Pollefeys M (2006) A general framework for motion segmentation: independent, articulated, rigid, non-rigid, degenerated and non-degenerate. In: 9th European Conference on Computer Vision. pp. 94–106. doi:https://doi.org/10.1007/11744085_8
34. Zhou F, Torre Fernando De L, Hodgins Jessica K (2008) Aligned cluster analysis for temporal segmentation of human motion. In: Automatic Face & Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on, pp. 1–7. doi:https://doi.org/10.1109/AFGR.2008.4813468
35. Zhou F, Torre Fernando De L, Hodgins Jessica K (2013) Hierarchical aligned cluster analysis for temporal clustering of human motion. IEEE Trans Pattern Anal Mach Intell 35(3):582–596